Evolutionary Algorithm for Decryption of Monoalphabetic Homophonic Substitution Ciphers Encoded as Constraint Satisfaction Problems

David Oranchak
NTU School of Engineering and Applied Science
Roanoke, VA 24018
doranchak@gmail.com

ABSTRACT

A homophonic substitution cipher maps each plaintext letter of a message to one or more ciphertext symbols [4]. Monoalphabetic homophonic ciphers do not allow ciphertext symbols to map to more than one plaintext letter. Homophonic ciphers conceal language statistics in the enciphered messages, making statistical-based attacks more difficult. We present a dictionary-based attack using a genetic algorithm that encodes solutions as plaintext word placements subjected to constraints imposed by the cipher symbols. We test the technique using a famous cipher (with a known solution) created by the Zodiac serial killer. We present several successful decryption attempts using dictionary sizes of up to 1,600 words.

Program Track: Real-World Applications

Categories and Subject Descriptors: E.3 Data Encryption: Code breaking

General Terms: Algorithms, Experimentation

Keywords: Evolutionary computing, genetic algorithms, Zodiac killer, Zodiac murder ciphers, codebreaking, cryptography, constraint satisfaction, homophonic substitution

1. INTRODUCTION

Simple substitution ciphers encrypt plaintext messages using symbols which map to individual plaintext letters. Monoalphabetic ciphers use the same mappings from plaintext to ciphertext throughout the encrypted message. Monoalphabetic substitution ciphers are often easy to decipher with frequency analysis because the simple mappings preserve letter frequencies of the plaintext message. Homophonic ciphers hide letter frequencies of plaintext messages. Each letter of enciphered plaintext is mapped to one or more ciphertext units, called *homophones*, which flattens the distribution of ciphertext symbols. The Zodiac killer is a famous serial killer who operated in California in the late 1960s [2]. In 1969, the killer sent three letters to area newspapers. In each letter, the killer took credit for recent shootings, and included a part of the 408-symbol three-part cipher (Figure 1[2]). A high school teacher and his wife soon decoded the cipher by hand.[2]: *[I like killing people because it is so much fun. It is more fun than killing wild game in the forrest because man is the most dangeroue animal of all to kill something gives me the most thrilling experence. It is even better than getting your rocks off with a girl. The best part is thae when I die I will be reborn in paradice and all the I have killed will become my slaves. I will not give you my name because you will try to sloi down or stop my collecting of slaves for my afterlife ebeorietemethhpiti.]* The killer mailed a second cipher to a San Francisco newspaper (Figure 1[2]). No satisfactory solution to this cipher has yet been found. We use the 408-symbol cipher as a test case for our technique, which we hope can be used to attack the 340-symbol cipher.

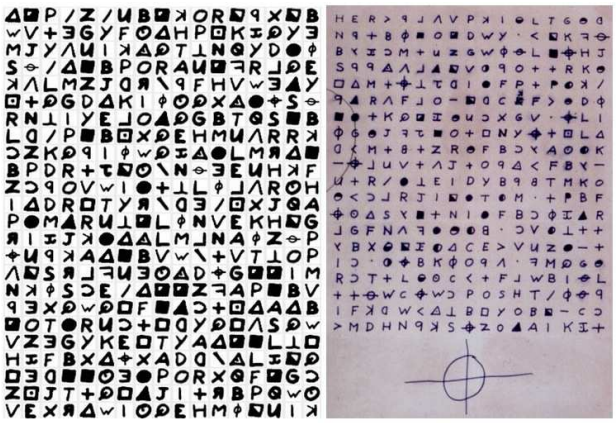

Figure 1: ***Left:*** **Solved 408-character homophonic substitution cipher sent by the Zodiac serial killer to three San Francisco newspapers.** ***Right:*** **Unsolved 340-character cipher sent by the killer.**

Many effective decryption techniques for simple ciphers have been studied, such as statistical analysis [3][1], evolutionary computing[8][6], and dictionary-based attacks [5][7]. Our experiments combine the strengths of evolutionary search and constraint-imposed dictionary-based attacks.

2. APPROACH

The 408-character cipher has a keyspace size of 26^{54}. To reduce the space, we attacked a 52-character region of ciphertext that decodes over 90% of the entire message. The targeted section decodes to the following: **killing wild game in the forrest because man is the most danger**. This section is only 12.7% of the cipher, but decodes 369 characters (90.4%) of the plaintext. We also limit word placements to sets of unique words having a minimum length

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-131-6/08/07.

Table 1: Results of experimental runs for different word pool sizes. F_e is the evolved solution's multi-objective fitness, and F_s is the multiobjective fitness of the known correct solution.

#Words	Correctness	F_e	F_s	Generations
500	1.0	58,880	58,880	874
859	1.0	82,1074	82,1074	668
1000	1.0	85,1099	85,1099	807
1395	1.0	104,1270	104,1270	1750
1600	0.9	110,1202	120,984	3222

of four. Our attack uses a variation of the techniques described by Olson[7] and Lucks[5]. We encode an attack as a non-conflicting set T of word and position selection tuples $[(w_i, p_i); 0 <= w_i < D, 0 <= p_i < 52$; D is dictionary size]. The GA performs generative placement of words into the 52-character cipher region. Words are selected from premade dictionaries composed of common words in the Zodiac corpus. The GA uses tournament selection of size two but sometimes selects individuals with lesser fitness. Crossover randomly merges feasible tuples from both parents. Simple mutation is applied to each offspring. All operators are restricted from producing infeasible tuple sets. Diversity is preserved using fitness sharing and elitism via Pareto sampling.

The first fitness measure is computed by counting partial or complete dictionary words that form beyond the 52-character region. The GA computes the second fitness measure by forming a graph G of the found words. Each word is a node, and any conflict between two words is an edge. We want to remove a minimal subset of nodes in G so no remaining edges (conflicts) remain, which is a minimum vertex cover problem. We obtain vertex cover within a factor of 2 of optimal by removing the maximal matching set of edges from the graph G to form G'. From G' the GA computes the second objective, Equation 1. Words from G' are counted for the entire cipher C except for the 52-character section (between k_0 and k_1). For each position, the score p_i is l (word length) if at least one word is found that covers position i, or 0 otherwise.

$$fitness_1 = \sum_{i \in C, i<k_0 \text{ or } i>k_1} \max_{l \in \{4,5,6,7,8,9,10\}} l \times p_i \qquad (1)$$

3. RESULTS

We ran experiments using the following parameters: population size 10,000; mutation probability: 0.1; genome size: 25 tuples; dictionary sizes: {500, 859, 1000, 1395, 1600}. Performance measure is determined by comparing the plaintext of the best evolved solutions to the plaintext of the known solution (using the given dictionary): **killing wild game inthe forrest because ??? isthe most danger** ("man" is missing because it is too short.) Results are shown in Table 3. Figure 2 plots performance versus number of generations for each dictionary size. Each experiment found a correct or very close to correct solution.

4. CONCLUSIONS

We encoded a homophonic substitution cipher attack as an evolutionary search of a combinatorial space of dictionary word placements subjected to constraints imposed by the ciphertext. By concentrating the search on a small 52-character section of the Zodiac killer's 408-character cipher, we reduced the search space and evolved correct decodings of the 52-character section that aid in the decryption of the entire ciphertext. Removal of approximate minimum vertex cover helps prevent exploration of word-dense plaintext decodings that in fact contain many conflicts.

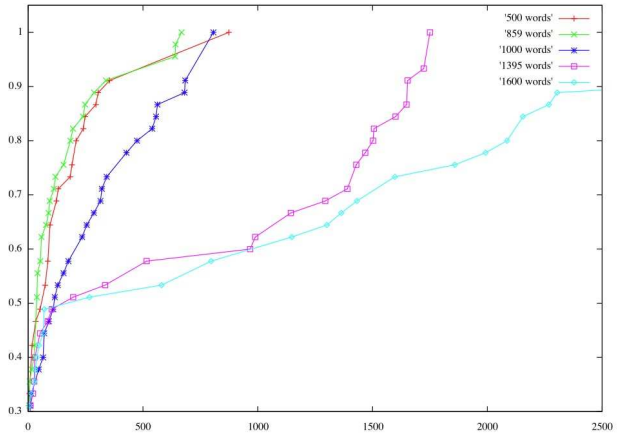

Figure 2: Performance plot for each dictionary size.

5. ACKNOWLEDGEMENTS

We thank the following people for their contributions and assistance: Chris McCubbin, Brax Sisco, and Edwin Olson.

6. REFERENCES

[1] J. M. Carrol and S. Martin. The automated cryptanalysis of substitution ciphers. *Cryptologia*, X(4):193–209, 1986.

[2] R. Graysmith. *Zodiac*. St. Martin's, New York NY, 1986.

[3] T. Jakobsen. A fast method for cryptanalysis of substitution ciphers. *Cryptologia*, 19(3):265–274, 1995.

[4] J. C. King and D. R. Bahler. A framework for the study of homophonic ciphers in classical encryption and genetic systems. *Cryptologia*, XVII(1):45–54, 1993.

[5] M. Lucks. A constraint satisfaction algorithm for the automated decryption of simple substitution ciphers. In *CRYPTO*, pages 132–144, 1988.

[6] R. A. J. Matthews. The use of genetic algorithms in cryptanalysis. 17(2):187–201, Apr. 1993. cryptanalysis; genetic algorithms; cryptographic systems; keyspaces; GENALYST.

[7] E. Olson. Robust dictionary attack of short simple substitution ciphers. 2007.

[8] R. Spillman, M. Janssen, B. Nelson, and M. Kepner. Use of a genetic algorithm in the cryptanalysis of simple substitution ciphers. 17(1):31–44, Jan. 1993.